DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 36, ARTICLE 2, PAGES 41–72
PUBLISHED 4 JANUARY 2017
http://www.demographic-research.org/Volumes/Vol36/2/
DOI: 10.4054/DemRes.2017.36.2

Research Article

Counting unreported abortions: A binomial-thinned zero-inflated Poisson model

Vidhura S. Tennekoon

Contents

Demographic Research: Volume 36, Article 2
Research Article

Counting unreported abortions: A binomial-thinned zero-inflated Poisson model

Vidhura S. Tennekoon[1]

Abstract

BACKGROUND
Self-reported counts of intentional abortions in demographic surveys are significantly lower than the actual counts. To estimate the extent of misreporting, previous research has required either a gold standard or a validation sample. However, in most cases, a gold standard or a validation sample is not available.

OBJECTIVE
Our main intention here is to show that a researcher has an alternative tool to estimate the extent of underreporting in a given dataset, particularly when neither a valid gold standard nor a validation sample is available.

METHODS
We adopt a binomial-thinned zero-inflated Poisson model and apply it to a sample dataset, the National Survey of Family Growth (NSFG), for which an alternative estimate of the average reporting rate (38%) is available. We show how this model could be used to estimate the reporting probabilities of intentional abortions by each individual in addition to the overall average reporting rate.

RESULTS
Our model estimates the average reporting rate in the NSFG during 2006–2013 as 35.3% (SE 8.2%). Individual reporting probabilities vary significantly.

CONCLUSIONS
Our estimate of the average reporting rate of the dataset used is qualitatively and statistically similar to the available alternative estimate.

CONTRIBUTION
The model we propose can be used to predict the reporting probability of abortions of each individual, which in turn can be used to correct the bias due to underreporting in

[1] Department of Economics, Indiana University Purdue University Indianapolis (IUPUI), School of Liberal Arts, Cavanaugh Hall Room 524, 425 University Boulevard, Indianapolis, IN 46202, USA.
E-Mail: vtenneko@iupui.edu.

any model in which the number of abortions is used as the dependent variable or as one of the covariates.

1. Introduction

In most demographic surveys, the number of intentional abortions is significantly underreported. For example, in the National Survey of Family Growth (NSFG), a survey of US women, only 38% of intentional abortions done during the years 2006-2013 are reported according to the estimates of the National Center for Health Statistics (NCHS). This estimate assumes that the data collected from abortion-service providers through a separate survey is accurate. Reporting rates of intentional abortions at the previous waves of the NSFG also have been estimated to be in the range of 35%–48%. Other major surveys of US women, such as the National Surveys of Young Women (NSYW) and the National Longitudinal Surveys of Work Experience of Youth (NLSY), too suffer from this limitation (Jones and Forrest 1992). Underreporting of this magnitude strictly limits the usefulness of self-reported abortion data. The NSFG warns that "NSFG data on abortion should not be used for substantive research" (USDHHS 2014: 35).

When intentional abortions are underreported, the total number of reported pregnancies is also automatically underreported, even if other pregnancy outcomes are accurately reported.[2] It hinders the accurate estimation of pregnancy timing, contraceptive efficiency, and many other impact-assessment studies in which the number of abortions or pregnancies could potentially be the dependent variable. Nevertheless, in many studies, self-reported abortion counts are analyzed to make strong conclusions while ignoring the impact of underreporting. For example, in Grindlay and Grossman (2013), cross-sectional data from the Department of Defense's survey of health-related behaviors among active-duty military personnel in 2005 and 2008 are analyzed to conclude that the rate of unwanted pregnancies (per 1000 women of reproductive age) has increased from 97 to 105 during the two years, notwithstanding the declared national goal of the government to reduce unwanted pregnancies. They also estimate that the unintended-pregnancy rate in the US military is 50% higher than that rate of the general US population, despite many formal and informal restrictions on sexual relationships faced by women (and men) in the military. Both of these

[2] The total number of abortions reported in 2006–2013 NSFG data is 3,288 while the total number of pregnancies is 30,035. If we assume that the reporting rate of intentional abortions is 38% the true number of abortions should be 8,653. If we also assume that the other pregnancy outcomes are reported correctly, the total number of pregnancies should be 35,400 which suggest that only 84.8% of pregnancies are reported.

observations could potentially be artifacts of differential reporting rates, across time periods in the first case and among the two groups in second case, due to the differences in the extent of perceived "social undesirability" of abortion. Another recent study, Tapales and Finer (2015), relies on the assumption that the reporting rates of intentional abortions of different subgroups do not vary across time periods.

Some researchers attempt to correct the bias due to underreported abortions before producing a related estimate. Lindberg (2011), another study on the unintended-pregnancy rate in the US military that analyzes data from the same survey as Grindlay and Grossman (2013), assumes that everyone in their sample reports only a fixed percentage of intentional abortions and therefore a fixed percentage of true pregnancies. The unreported pregnancies are assumed to be 11.9% of the number of reported cases based on two previous studies that estimate the reporting rates of intentional abortions in the NSFG in comparison to provider surveys. In addition to ignoring the differences in reporting rates across various subgroups, this study simply ignores that the average reporting rate could vary significantly across different surveys.

If the reporting rates of intentional abortions are known, preferably at the individual level, it is possible to correct for any potential bias due to underreporting in many studies. When the reporting rate at a particular survey is not known, researchers in most cases simply ignore misreporting, as in Grindlay and Grossman (2013), or use an estimate from a previous study, which may or may not be based on the same survey, as in Lindberg (2011). The estimation of reporting rates in a given dataset usually requires a reliable external source that can be treated as a gold standard. Data collected from abortion-service providers (Jones and Forrest 1992; Fu et al. 1998; Finer and Zolna 2014), Medicaid claims (Jagannathan 2001), and medical records (Udry et al. 1996) have been used for comparisons, ignoring that these reference datasets too could be vulnerable to underreporting (Henshaw 1998). Some researchers have attempted to estimate underreported abortion counts without assuming a gold standard (Yan, Kreuter, and Tourangeau 2012, for example), but their approaches still require a validation sample. Researchers have attempted to estimate underreported count data using validation samples in various other settings too. In Amoros, Martin, and Laumon (2006) a validation sample is used to estimate the extent of underreporting of road-crash casualties in France. Dvorzak and Wagner (2016) use a binomially-thinned Poisson model to capture the extent of underreporting in the number of cervical-cancer deaths and use a validation sample to identify the model through a Bayesian approach.

The challenges faced by demographers due to underreported count data are documented in several studies. Some examples are incidents of sexual violence (Gross et al. 2006), concurrent sexual partnerships (Adimora, Schoenbach, and Doherty 2007), and birth counts in Chinese census data, which lead to underestimated fertility rates (Goodkind 2004) and implausible sex ratios (Goodkind 2011). Among the other areas where underreported counts have been found and a model similar to the one we use could be used are in estimating the frequency of absenteeism in workplaces, the reporting of industrial injuries, the violations of safety regulations in nuclear power plants, criminal victimizations, needlestick injuries at hospitals, and earthquakes and cyclones (see Cameron and Trivedi 2013: 488 for details).

The binomially-thinned zero-inflated Poisson model we adopt here can be used to estimate not only the average reporting rates but also the reporting probabilities of each individual without using any gold standard or a validation sample. As we show here, this model, which combines the features of the zero-inflated Poisson model (Lambert 1992) and the binomial-thinned Poisson model (Winkelmann and Zimmermann 1993), predicts the average reporting rate of abortion counts in NSFG data with remarkable closeness to the available estimates. Cameron and Trivedi (2013) has a detailed discussion about these models as well as their applications.

Our model is closely related to the binomially-thinned zero-inflated negative binomial model (ZI-NB2-Logit model) proposed by Papadopoulos (2014) to estimate underreported crime self-reports in England and Wales. They generalize the binomially-thinned negative binomial model (NB2-Logit model, using their terminology) developed in Winkelmann and Zimmermann (1993) by allowing zero-inflation.

The method we propose can be used in many related studies where the accuracy of an estimate is jeopardized by the underreported abortion counts. The model can be used directly when the interest is on the impact of some factor in determining the number of intentional abortions, or the number of pregnancies, by simply including a variable capturing that factor as a covariate. When the abortion count should enter as a covariate, the estimated reporting rates can be used to correct for any bias due to underreporting.[3]

In section two, we discuss other studies that estimate reporting rates of abortion rates in various surveys. In section three, we present our model. Our data is explained in section four, and the results are presented and analyzed in section five. Section six concludes.

2. Reporting rates of abortion counts in various surveys

Various researchers have attempted to estimate the extent of underreporting of abortion counts in surveys of US women at least since Jones and Forrest (1992), which is a detailed study that covers multiple waves of three major surveys: the 1976, 1982, and 1988 cycles of the NSFG; the 1976 and 1979 NSYW; and the NLSY. Their estimations of reporting rates are based on external data collected by the Alan Guttmacher Institute

[3] When the estimated reporting rates are used in this manner in an econometric model, the additional randomness introduced by these estimates should be reflected in the standard errors.

(AGI) from abortion-service providers. While the NSFG dataset includes all pregnancies of the respondents aged 15–44, the Jones and Forrest (1992) analysis is limited to the pregnancies during the most recent years (3–4 years, depending on the NSFG cycle), for which more-accurate data is available. The estimates of reporting rates during each of the 11 calendar years covered in the three NSFG cycles vary from 25% to 60%, while the average reporting rate during all these years was 35%. Reporting rates in the two waves of NSYW, which covers a sample of adolescents, have been estimated at 59% and 42%, respectively. The overall reporting rate in NLSY has been estimated at 40%. The authors caution the readers about the measurement error ignored in their "gold standard," which potentially causes these estimates to be upward biased.

Fu and colleagues (1998) extend the analysis of Jones and Forrest (1992) to the 1995 cycle of the NSFG following the same methodology. During the four years that this cycle covers, the reporting rate is estimated to be within 42%–47%. This cycle of the NSFG also introduced an audio computer-assisted, self-administered survey to improve the reporting accuracy of sensitive topics. Fu and colleagues (1998) show that the reporting rates increased to 57%–60% when this procedure was used. The next study of this series, Jones and Kost (2007), repeat the same procedure and find that the reporting rate in 2002 cycle of the NSFG is 47%.

The studies comparing survey responses with provider data are the main sources of reporting-rate estimates of abortion counts in US demographic surveys. However, these are likely to be overestimates, as the provider data they use for comparison also suffers from underreporting problems. As Henshaw (1998) estimates, the reporting rate of this provider data in 1996 is only 96%–97%, which suggests that the reporting rates estimated based on the provider data could potentially be 3%–4% higher than the actual rates. This finding emphases the need for a reliable gold standard to compare the accuracy of self-reported abortion counts with.

Some researchers have used medical records and insurance-claim data for comparison, but these studies are based on relatively small samples. Udry and colleagues (1996), for example, compare self-reported abortion counts of 104 women aged 27–30 with their medical records to find that 19% of them do not report one or more of their abortions. Jagannathan (2001) uses Medicaid claim data as her gold standard to find that only 29% of abortions are self-reported. Her study is based on a sample of 1,236 mothers on welfare. Yan, Kreuter, and Tourangeau (2012), the only team of researchers who use a model-based approach to estimate the average reporting error of abortion counts in the NSFG, conclude that their technique fails to correctly estimate the average reporting rate of abortions after finding an implausibly low error rate in the range of 1%–3%.

While the reporting rate of abortion counts in major US surveys, as estimated in most previous research, is within 25%–60%, there is significant variation across different surveys and different cycles of the same survey. The variation in reporting rates across different demographic groups is even larger. The reporting rate of Whites was higher than the reporting rate of non-Whites in all datasets analyzed by previous researchers. Underreporting rates were relatively lower for married women. These comparisons, however, are based on the unconditional means and may illustrate the impact of one or more other factors. The impact of other variables on the reporting probability is ambiguous. In the 1982 NSFG dataset, the reporting rate increases with age, but in the 1988 dataset, there is no clear pattern (Jones and Forrest 1992). As the same study shows, the reporting rate decreases with age in the NLSY dataset. Two other studies (Fu et al. 1998; Jones and Kost 2007) show a U-shaped relationship between the age and the reporting rate. So not only do we notice variation in reporting rates across different demographic groups in a given dataset, but we also observe differences in this pattern of variation across different datasets. Therefore, making inferences about the reporting rate of abortions by respondents in one dataset using the estimates from another does not guarantee a better result and may even lead to a worse outcome.

The reporting rate of White women, as shown in all previous research, is consistently higher than the reporting rate of non-White women. Black women in particular have a lower reporting rate. In addition, the reporting rate of Hispanic women is significantly lower than the reporting rate of non-Hispanic women. Though the causality of the relationship between the racial identification of a woman and the reporting rate of abortions is not obvious, this empirical observation has been reported consistently by several researchers. Since we use this relationship to impose sign restrictions when identifying our model parameters, the relevant empirical observations are detailed below.

As estimated by Jones and Forest (1992), the reporting rate of White women in the 1982 NSFG is 21% higher than the rate of non-White women. In the 1998 NSFG, the reporting rate of White women is 38%, while it is only 27% for non-White women. Fu and colleagues (1998) shows that the reporting rate of White women is 12%–17% higher than the rate of non-White women. In Jones and Kost (2007), the estimated reporting rate of Hispanic women is only 29%, a statistically different rate from the number for non-Hispanic White women (61%). The reporting rate for non-Hispanic Black women (42%) too is significantly below the rate for non-Hispanic White women. A multivariate logistic regression analysis of Udry and colleagues (1996) demonstrates that, out of the variables that they examined, race was the strongest predictor of abortion underreporting. Based on their analyses, non-White women are 3.3 times as likely as Whites to underreport their intentional abortions. This view is also confirmed by Jagannathan (2001). In all four of her logistic regression models, the coefficients of Black and Hispanic variables are significantly negative against the excluded Whites.

Jagannathan (2001) suggests three possible reasons for the observed relationship between the racial identification of a woman and her reporting rate of abortions:

a) The interaction of race and abortion attitudes,
b) Interracial differences in other attitudes that affect underreporting, which are not controlled for, and
c) Interracial differences in the degree of mistrust on the survey process.

Based on the above arguments and empirical findings, we assume that the coefficients of Black and Hispanic dummy variables in our model should be negative.

3. The model

3.1 Poisson models

The Poisson regression model, derived from the single-parameter Poisson probability distribution, is the most basic option for modelling a count variable (Y), which can only take a nonnegative integer value (y). It is a single index model in which the covariates (X) are usually linked with the Poisson mean (λt) using an exponential function to impose non-negativity:

$$\Pr(Y = y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!} \text{ where } \lambda = \exp(X\beta). \tag{1}$$

The two terms λ and t represent the mean number of counts during a unit period of exposure and the length of exposure, respectively, while β is a vector of parameters that can be estimated using MLE or another suitable estimation technique when Y can be observed without any measurement error. In most applied work, the exposure period does not vary across observations and therefore is normalized to unity. The exposure term does not appear explicitly in those models. In our application, t shows significant variation, and we use this variation to strengthen the identification of our model.

There are some well-known limitations of the above basic model. The first is its equi-dispersion assumption, the variance being equal to the mean. In practice, most real data shows over-dispersion, a variance larger than the mean. The negative binomial model is a generalization of the basic Poisson model that includes one additional over-dispersion parameter. The second issue is that many real-count variables have more zeros than the number produced by a Poisson process, conditional on its mean. When a part of these zeros is excluded, these distributions better represent a Poisson distribution. This has motivated several variants of zero-inflated Poisson models. We focus on the two-part model presented in Lambert (1992), which assumes that the final observed outcome is the result of a two-staged process.

The first stage of the Lambert (1992) model (not following his notation) is a Bernoulli zero-generation process where the outcome is equal to 0 with probability φ. When the outcome of this Bernoulli process is 1, which happens with probability $(1-\varphi)$, it triggers a Poisson random-number generation process to produce a nonnegative integer outcome. The final observed outcome includes zeros generated through both mechanisms and positive integers generated by the Poisson process.

$$\Pr(Y = y|y > 0) = (1-\varphi)\frac{(\lambda t)^y e^{-\lambda t}}{y!} \tag{2}$$

$$\Pr(Y = 0) = \varphi + (1-\varphi)e^{-\lambda t} \tag{3}$$

The Poisson mean is linked with the covariate vector (X) using an exponential function, while the Bernoulli mean is linked with the covariates (Z) using an inverse logit function such that $\lambda = \exp(X\beta)$ and $\varphi = \frac{\exp(Z\delta)}{1+\exp(Z\delta)}$, where δ is the vector of coefficients of the zero-inflation process. The two covariate vectors, X and Z, can be identical, overlapping, or disjointed. This model takes account of the existence of both excess zeros and over-dispersion.

Underreporting is one of the mechanisms that generates excess zeros. If the only way that an observation could be underreported is through a Bernoulli process that randomly assigns a 0 to a true positive value, we can use the above zero-inflated Poisson (ZIP) model to fully explain the underreporting mechanism. In the case of self-reported counts of an outcome, this requires each respondent to either report her count with perfect accuracy or report a zero count irrespective of her true count. In reality, however, respondents may report a number lower than their true counts, but not always a zero. Therefore, misreporting has to be modeled in a different manner. The binomial thinning process closely mimics the true underreporting process of count data (Winkelmann and Zimmermann 1993) and assumes that the observed outcome is a "Poisson-stopped sum of Bernoulli random variables" (Fader and Hardie 2000). More specifically, if the true variable, generated through a Poisson process, is Y^* and the observed variable is Y, they are related as $Y = \sum_{i=1}^{Y^*} I_i$ where I_i is a Bernoulli random variable that takes the value of 1 with probability π. The parameter π can be interpreted as the reporting rate. This relationship is expressed using the binomial thinning operator $\circ$ as

$$Y = \pi \circ Y^* = \pi \circ \frac{(\lambda t)^y e^{-\lambda t}}{y!} \tag{4}$$

It can be shown that the binomial thinning of a Poisson distributed random variable produces another Poisson distributed random variable such that

$$\pi \circ Poisson(\lambda t) = Poisson(\pi \lambda t). \tag{5}$$

In a regression framework, the most commonly used “Poisson logit model” (Winkelmann and Zimmermann 1993) links π with covariates using an inverse logit function while the Poisson mean, λt, is linked using an exponential function as in (1).

3.2 Underreported abortions

The outcome we are interested in is the total number of intentional abortions done by women in their reproductive age. We have these numbers reported by a cross-sectional sample of US women aged 15–45 years but with potential measurement error. In this sample, an observed zero count could mean one of several things. First, if a woman never had sex, there’s no way that she could get pregnant (we ignore the possibility of artificial insemination by a woman who never had sex). They have zero counts of intentional abortions as well as zero counts of other types of pregnancies. Second, not all women who have sex get pregnant; some have fertility issues, and some do not want to get pregnant for various social, economic, medical, and personal reasons. They avoid pregnancies through contraceptives. This second group also has zero intentional abortions and zero pregnancies. Women in the third group have sex and get pregnant, but an intentional abortion is never a choice for them because either they have social, religious, or ethical concerns or they well plan their pregnancies. They also have zero intentional abortions but nonnegative counts of pregnancies. The last group does have some intentional abortions but they report their abortion counts as zeros. If a woman in this fourth category doesn’t have any pregnancy outcome except these unreported intentional abortions, her total pregnancy count too is incorrectly recorded as zero. At the same time, a reported positive abortion count indicates that the woman has had sex, been pregnant, and had some abortions. However, the actual count of abortions could be higher than the number reported.

If we can correctly identify the subset of women who have ever been pregnant, we can isolate the first two groups and analyze the rest of the observations. But when the number of intentional abortions is inaccurate, so is the total number of pregnancies. Therefore, we have no way to isolate the second group. However, if all self-reported virgins are truthful, we can isolate the first group. While self-reported virginity too is prone to misreporting (Tennekoon and Rosenman 2014), we do not expect that to be a serious issue in this sample. On that basis, we assume that the true number of pregnancies, and therefore the true number of intentional abortions, is zero if a woman reports that she never had vaginal intercourse.[4]

Our main target is to estimate the reporting rate of intentional abortions. A woman is exposed to pregnancy from the first day that she has vaginal intercourse or her menarche, whichever happens later. The window that she is exposed to an intentional abortion begins approximately a month later. Ignoring that relatively short delay, we measure exposure from the date that she first had sex after her menarche until the date she was interviewed (t). If $t = 0$, there is no way that a woman could have any pregnancy, and therefore, the count of intentional abortions should be zero. The actual number of intentional abortions of a woman (i) that we do not observe (Y_i^*), assumed to be determined by a ZIP process as in Lambert (1992), is a function of her exposure (t_i); the probability of being in the group of women that includes those who decide not to be pregnant, are infertile, or are strictly against abortion (φ_i); and the expected number of intentional abortions during a unit period (λ_i). If the probability of reporting each intentional abortion by a woman is π_i, the observed outcome can be expressed as

$$\Pr(Y_i = y_i | y_i > 0) = (1 - \varphi_i) \frac{(\pi_i \lambda_i t_i)^{y_i} e^{-\pi_i \lambda_i t_i}}{y_i!}; \tag{6}$$

and

$$\Pr(Y_i = 0) = \varphi_i + (1 - \varphi_i) e^{-\pi_i \lambda_i t_i}. \tag{7}$$

The log-likelihood function takes the following form:

$$LnL = \sum_{Y_i=0} \ln[\varphi_i + (1 - \varphi_i) \exp(-\pi_i \lambda_i t_i)] + \sum_{Y_i>0} [\ln(1 - \varphi_i) - \pi_i \lambda_i t_i + Y_i \ln(\pi_i \lambda_i t_i) - \ln(Y_i!)]. \tag{8}$$

In a regression framework, each of the two Bernoulli means are linked with the covariates using the inverse logit function, and the Poisson mean is linked using the exponential function, as done usually, so that $\varphi_i = exp(Z_i\delta)/(1 + exp(Z_i\delta)), \pi_i = exp(R_i\gamma)/(1 + exp(R_i\gamma))$ and $\lambda_i = exp(X_i\beta)$. We finally have a triple-index model with three vectors of parameters, δ, γ , and β, to be estimated. The exposure variable, t_i, acts as an offset parameter on the index $X_i\beta$. This is because $\lambda_i t_i = exp(X_i\beta + \ln(t_i))$. It can easily be handled within the log-likelihood function in (8) by including

[4] A robustness check to find the sensitivity of our results to any measurement error in self-reported virginity that we present and discuss later supports this assumption.

$\ln(t_i)$ as an additional covariate of vector X_i, of which the coefficient is restricted to 1. If $t_i = 0$ for any observation, that observation does not contribute to the value of the above log-likelihood function. Therefore, ignoring the observations with zero exposure does not affect the maximum likelihood solution.

This model is identified using the variation in π_i and cannot be identified when π_i is a constant. This is because $\ln(\pi_i)$ is indistinguishable from the constant term in the linear index $X_i\beta$ (Cameron and Trivedi 2013). Even when π_i has sufficient variability, we face another subtle identification issue (Papadopoulos and Santos Silva 2012). That is, the parameter vector $[\beta, \gamma]$ is indistinguishable from the parameter vector $[\beta + \gamma, -\gamma]$ when the same regressors are used in both processes (i.e., $X_i = R_i$) since

$$\frac{exp(R_i\gamma)}{1+exp(R_i\gamma)} exp\ (X_i\beta) = \frac{exp(-R_i\gamma)}{1+exp(-R_i\gamma)} exp\ (R_i\gamma + X_i\beta). \tag{9}$$

This issue can be avoided by sign restrictions on at least one of the covariates in R_i based on prior information or by including at least one variable that does not belong to X_i in vector R_i. Even when R_i includes one or more variable that does not belong to X_i, there are two maxima, but the global maximum is distinguishable from the other since the log-likelihood values are different (Staub and Winkelmann 2013).

In our application, π_i is well known for its variation, and we also have prior information to impose some sign restrictions as explained in the previous section. Therefore, we do not need to rely on functional form assumptions alone for identifying our model. Once the parameters of the model are estimated using maximum likelihood, the average reporting rate of intentional abortions in this dataset can be estimated as $E(\hat{\pi}_i) = \frac{1}{N}\sum_{i=1}^{N} \exp(R_i\hat{\gamma})/(1 + exp(R_i\hat{\gamma}))$. The average reporting rate of a given subgroup (for comparison with previous research) can be calculated by averaging the predicted reporting rates for observations belonging to that subgroup, in a similar manner. This is because

$$E(\hat{\pi}_i|G_i = 1) = \frac{E(G_i\hat{\pi}_i|G_i=1)}{E(G_i)} = \frac{\sum_{i=1}^{N} G_i[\exp(R_i\hat{\gamma})/(1+exp(R_i\hat{\gamma}))]}{\sum_{i=1}^{N} G_i} \tag{10}$$
$$= \frac{1}{N_G}\sum_{i=1}^{N_G} \frac{exp(R_i\hat{\gamma})}{(1+exp(R_i\hat{\gamma}))}$$

Here, G_i is a dichotomous variable which takes the value 1 for all women belonging to a given subgroup and 0 otherwise. N_G is the total number of women in that subgroup.

4. Data and estimation

For this study, we use data from the two most recent cycles of the NSFG, which is a cross-sectional survey of men and women aged 15-44 years who are US residents. It has collected information on family life, marriage and divorce, pregnancy, infertility, the use of contraception, and men's and women's health since 1973. Our data sample covers the most recent data releases on female respondents (as at January 1, 2015), which includes the survey years 2006–2013. The 2006–2010 release surveys 12,279 respondents, while the most recent 2011–2013 release collects data from additional 5,601 women. The two data releases of this multi-stage probability-based sample survey are not different in their structure or survey design. In both cycles, Blacks, Hispanics and teens are oversampled. The reporting rate of abortions in each of these cycles also has been estimated to be equal (USDHHS 2014). The interviews have been conducted by trained female interviewers using laptop computers. We only use the variables available in public-use files at no cost to the researcher.

Out of the total of 17,813 respondents in our complete dataset, 2,393 (13.4%) report never having sex. Our focus in this study is the remaining 15,420 women who have had sex. Among them, 4,394 (28.5%) have never been pregnant, while others report 1 to 20 pregnancies. Only less than 15% of these women self-report ever having an intentional abortion. The number of reported abortions ranges from 1 to 16. The exposure variable combines information contained in three other variables. First, the start of the exposure window was calculated as the first time a woman had sex or her menarche, whichever happened later. Exposure is the number of years elapsed from that point until the survey was done.

Most of the variables in this dataset represent the status of these women when they were surveyed. For example, income, marital status, and religious affiliation show their current status. The number of abortions, collected retrospectively, is related to the decisions they made in the past, and the current status may not explain why they made a decision several years ago. Therefore, we have to choose our variables carefully. Also, most of the variables that would potentially cause a woman to be in the zero-abortions group are also likely to affect their number of intentional abortions if they belong to the potential-abortions group. We include age (as a categorical variable), race, Hispanic origin, years of education, a dummy for having some college education (years of education being more than 12 years), religious affiliation, sexual orientation, number of formal marriages, number of cohabiting unions, a dummy for ever being employed, a dummy for ever seeking help to get pregnant, a dummy for ever receiving treatment for infertility, a dummy for ever using a contraceptive, a dummy that is equal to 1 if the respondent has never had a job, and two additional dummies to capture cohort effects (born in 1980s and 1990s) as the covariates of the equation that explains excess zeros.

These variables capture the factors causing a woman not to be pregnant and also the factors that motivate her not to abort a pregnancy if she gets pregnant. The current religious affiliation and the sexual orientation may have changed during a woman's lifetime prior to the survey, but we assume those changes to be low-frequency events.

Most of these variables are likely to affect the number of abortions too, if a woman gets pregnant. Therefore, all these variables except the two dummies identifying the birth cohort enter the equation explaining the Poisson mean. In addition, we include the number of live births and the number of times an emergency contraceptive pill was used in that equation. The exposure variable also enters this equation in its log form with its coefficient restricted to 1. Unlike the decision to get pregnant or have an abortion, the decision to report an abortion is a contemporary decision. Our main interest is in estimating the parameters of this equation. We include the variables identified by previous researchers as affecting the reporting probability of abortions in this equation. They include race, Hispanic origin, the age categories we used in the other two equations, the number of years of education, a dummy for having some college education, marital status, religious affiliation, and income categories. Summary statistics of these variables (unweighted) are presented in Tables 1 and 2.

Table 1: Summary statistics for continuous and count variables

Variable	Mean	Std. Dev.	Weighted Mean	Weighted Std. Dev
Number of abortions reported	0.21	0.63	0.19	0.61
Age (years)	30.12	7.81	31.11	7.98
Education (Number of years)	13.19	2.66	13.50	2.69
Number of live births	1.38	1.41	1.44	1.44
Number of emergency contraceptives	0.28	1.22	0.26	1.57
Number of cohabiting unions	0.94	1.23	0.94	1.59
Exposure (Number of years)	13.97	7.71	14.82	7.92

The parameters of the model were estimated using the "ml" command in Stata (StataCorp 2015), and the reporting rates were calculated using Mata. A sample Stata/Mata code that explains the estimation procedure using simulated data is given in the appendix. The data extract used to produce the empirical results and the Stata/Mata code is available as an electronic supplement. Since the NSFG is a multi-stage probability-based sample survey, sample weights were used in all estimations. Standard errors were calculated using both the empirical Hessian estimator and the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

Table 2: Summary statistics for categorical variables

Variable	Mean	Weighted Mean
Age		
15–19 years	0.10	0.08
20–24 years	0.18	0.17
25–29 years	0.21	0.19
30–34 years	0.19	0.17
35–45 years	0.32	0.39
Birth cohort		
Born in 1980s	0.38	0.35
Born in 1990s	0.10	0.09
Race		
White	0.65	0.73
Black	0.23	0.16
Other	0.12	0.11
Hispanic origin	0.23	0.18
Religion		
None	0.20	0.20
Catholic	0.24	0.24
Protestant	0.48	0.48
Other	0.07	0.08
Marital status		
Never married	0.37	0.29
Married	0.37	0.46
Previously married	0.12	0.10
Cohabiting	0.14	0.14
Poverty level income		
Below 100%	0.29	0.24
100–199%	0.24	0.23
200–299%	0.17	0.17
300–399%	0.13	0.16
400–499%	0.08	0.10
Above 500%	0.09	0.10
Education		
More than 12 years	0.52	0.57
Sexual orientation		
Heterosexual	0.92	0.98
Homosexual	0.01	0.00
Bisexual	0.07	0.02
Pregnancy intention		
Never used contraceptive	0.01	0.01
Ever sought help to get pregnant	0.08	0.09
Ever used infertility services	0.12	0.13
Never worked	0.02	0.02

5. Results

5.1 The true count of abortions

The coefficient estimates of the factors affecting the true count of abortions are reported in Tables 3 and 4. Younger cohorts are more likely to be in the zero-abortions group than the older ones. Since the women born in the 1990s and 1980s are less likely to abstain from sex compared to the older cohorts, this may reflect their increased knowledge about reproductive health, their awareness about contraceptives in particular. It may also show the increased ability of women to negotiate with their sexual partners, which in turn helps in avoiding unintended pregnancies.

Table 3: Excess zeros

Variable	Estimate	Std. Err. I		Std. Err. II	
Age (Baseline: 35–45 years)					
15–19 years	–1.432	0.785	*	1.103	
20–24 years	–0.014	0.283		0.436	
25–29 years	–0.174	0.232		0.328	
30–34 years	0.262	0.155	*	0.270	
Birth cohort					
Born in 1980s	0.304	0.164	*	0.223	
Born in 1990s	1.281	0.355	***	0.427	***
Hispanic origin	0.024	0.171		0.280	
Race (Baseline: White)					
Black	–0.860	0.160	***	0.265	**
Other	–0.213	0.224		0.453	
Education					
Education (years)	0.079	0.039	**	0.062	
More than 12 years	–0.940	0.189	***	0.312	***
Religion (Baseline: No religion)					
Catholic	0.406	0.210	*	0.351	
Protestant	0.819	0.164	***	0.274	***
Other	0.928	0.211	***	0.356	***
Pregnancy intention					
Never used contraceptive	–5.892	27.113		4.931	
Ever sought help to get pregnant	0.380	0.370		0.603	
Ever used infertility services	–0.089	0.292		0.492	
Sexual orientation (Baseline: Heterosexual)					
Homosexual	–0.676	2.552		2.812	
Bisexual	–0.734	0.495		0.610	
Number of formal marriages	–0.951	0.064	***	0.109	***
Number of cohabiting unions	0.425	0.096	***	0.178	**
Never worked	0.940	0.418	**	1.000	
Constant	–0.058	0.525		0.793	
Number of observations	15,401				
Log likelihood	–6958.20				

* p<0.1,**p<0.05,***p<0.01
Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

Within each (decennial) cohort, age increases the probability of being in the zero-abortions group, and it also increases the count of abortions. In other words, the fraction of women with zero abortions among the younger age groups is smaller than among the older, but the mean number of abortions of those younger women who would consider having an abortion during a given period is smaller compared to older women. This suggests some selection behavior as these women age; those who abstain from intentional abortions during the early ages continue to do so, while those who had any abortions when young have more.

Table 4: The count of abortions

Variable	Estimate	Std. Err. I		Std. Err. II	
Age (Baseline: 35–45 years)					
15–19 years	–1.924	0.709	***	1.140	*
20–24 years	–0.567	0.361		0.617	
25–29 years	–0.525	0.280	*	0.574	
30–34 years	0.400	0.251		0.477	
Number of live births	0.017	0.018		0.026	
Number of emergency contraceptives	0.040	0.003	***	0.005	***
Hispanic origin	1.035	0.388	***	0.660	
Race (Baseline: White)					
Black	1.083	0.329	***	0.576	*
Other	0.204	0.281		0.672	
Education					
Education (years)	–0.196	0.058	***	0.088	**
Education: More than 12 years	–0.744	0.426	*	1.139	
Religion (Baseline: No religion)					
Catholic	–0.728	0.236	***	0.371	**
Protestant	0.100	0.215		0.339	
Other	–0.350	0.299		0.487	
Pregnancy intention					
Never used contraceptive	–1.941	0.532	***	0.655	***
Ever sought help to get pregnant	–0.179	0.209		0.307	
Ever used infertility services	–0.174	0.160		0.193	
Sexual orientation (Baseline: Heterosexual)					
Homosexual	–0.849	0.983		1.094	
Bisexual	–0.287	0.181		0.190	
Number of formal marriages	–0.002	0.010		0.012	
Number of cohabiting unions	0.088	0.056		0.117	
Never worked	0.699	0.276	**	0.636	
Constant	0.894	1.097		1.759	
Ln(Exposure): Restricted	1.000				
Number of observations	15,401				
Log likelihood	–6958.20				

* p<0.1,**p<0.05,***p<0.01

Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

The total number of emergency contraceptives (such as morning-after pills) has a positive effect on the count of abortions. This variable is a proxy for the tendency for unsafe sex, which increases the chances of intentional abortions by a woman. The coefficient of the total number of live births, used to measure the marginal value of an additional child, has the positive sign as expected but is statistically insignificant. Since these two variables explain the characteristics of a woman who would potentially have an abortion rather than the selection to the same group, they were only used in the equation explaining the mean count of abortions.

Blacks are less likely to be in the zero-abortion group and are also likely to have more intentional abortions than Whites. Those belonging to other race categories than Blacks and Whites also show a behavior similar to Blacks, but the magnitudes are not as high. Hispanics do not differ significantly from non-Hispanics in their probability to be in the zero-abortion group, but their mean number of abortions is larger. The coefficients for being in the zero-abortion group are positive for all religious categories (Catholic, Protestant, and other) compared to the group without a religious identity. The coefficients of Catholics and the other religious categories are negative on the count of true abortions, but only the coefficient for being Catholic is statistically significant. The coefficient for being Protestant is positive but small and statistically insignificant.

Education negatively affects the mean number of abortions, probably showing that educated women are more likely to plan their pregnancies. The effect is stronger for women with some college education. However, those with a college education are less likely to be a part of the zero-abortion group, probably showing their increased exposure to sexual activity. Women who were never employed, on the other hand, have positive and significant effects on both outcomes. Perhaps this group is less exposed to sex and therefore to pregnancy. However, if they are exposed, they have less incentive to have an abortion since their cost of having an additional child is not as high as the same cost for a working woman.

Those who never used a contraceptive tend to have a lower number of abortions. This group may include women who do not want to interfere with the natural processes of pregnancy and childbirth due to religious or ethical concerns, making them less likely to have an abortion, as well as some other women who want to have more children simply because their utility from an additional child does not diminish substantially after having a few children. Both coefficients for being homosexual and bisexual were insignificant, perhaps due to a lack of variation in data. The number of cohabiting unions increases the probability of being in the zero-abortion group. The number of formal marriages, however, decreases the chance of zero abortions. The coefficients of the usage of infertility services and the usage of any help to get pregnant were both insignificant.

5.2 The reporting rates

The focus of our study is the reporting rate of abortions and the factors affecting that rate, not the factors affecting the true number of abortions. In Table 5, we report the coefficient estimates of the factors potentially affecting the reporting rate. The reporting rate is relatively high for the younger age groups and drops as these women become older. Blacks have lower reporting rates than Whites, and Hispanics have lower rates than non-Hispanics. Other racial groups also have lower reporting rates than Whites. The reporting rate increases with the level of education, with a step increase when a woman has some college education. Married, previously married (divorced, separated, or widowed), and cohabiting women are all less likely to report abortions than never-married women. Catholics are more likely to report an abortion than a person with no religious affiliation, but Protestants and those in other religious categories do not differ significantly in their reporting rates of abortions than non-religious women. The effect of income on reporting is somewhat ambiguous. In general, those in middle income categories are more likely to report abortions than those in lower and higher income categories.

Table 5: Reporting of abortions

Variable	Estimate	Std. Err. I		Std. Err. II	
Age (Baseline: 35–45 years)					
15–19 years	2.454	0.942	***	1.565	
20–24 years	1.500	0.420	***	0.802	*
25–29 years	0.897	0.345	***	0.792	
30–34 years	–0.252	0.301		0.539	
Hispanic origin	–1.194	0.437	***	0.786	
Race (Baseline: White)					
Black	–1.064	0.357	***	0.665	
Other	–0.238	0.361		0.826	
Education					
Education (years)	0.330	0.063	***	0.104	***
Education: More than 12 years	0.462	0.520		1.357	
Religion (Baseline: No religion)					
Catholic	0.692	0.309	**	0.531	
Protestant	–0.164	0.266		0.408	
Other	–1.475	0.373	***	0.668	**
Marital status (Baseline: Never married)					
Married	–0.598	0.110	***	0.164	***
Previously married	–0.256	0.111	**	0.175	
Cohabiting	–0.362	0.099	***	0.146	**
Income (Baseline: Below 100%)					
100–199%	–0.009	0.079		0.110	
200–299%	0.205	0.091	**	0.137	
300–399%	0.192	0.106	*	0.157	
400–499%	0.284	0.126	**	0.206	
Above 500%	–0.039	0.146		0.284	
Constant	–5.586	1.073	***	1.204	***
Number of observations	15,401				
Log likelihood	–6958.20				

* p<0.1,**p<0.05,***p<0.01

Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

Overall, the reporting rate of abortions by the respondents of the NSFG during the years 2006–2013 is estimated to be 35.28% by our baseline model. While the average reporting rate is at this level, the estimated reporting rate varies significantly across individuals. In Figure 1, we show the distribution of the estimated reporting probability of different age groups and for the full sample. In Table 6, we present estimated reporting rates of different demographic groups together with the standard errors linearized using the delta method. The estimated average reporting rate from our baseline model is slightly lower than the estimate of USDHH based on comparisons with provider data but statistically indistinguishable from their estimate of 38% with the level of precision we have. Our estimate makes more sense when we adjust USDHH estimates by eliminating the upward bias due to 3%–4% underreporting in provider data (Henshaw 1998).

Figure 1: Distribution of estimated reporting probabilities within different age groups and in full sample

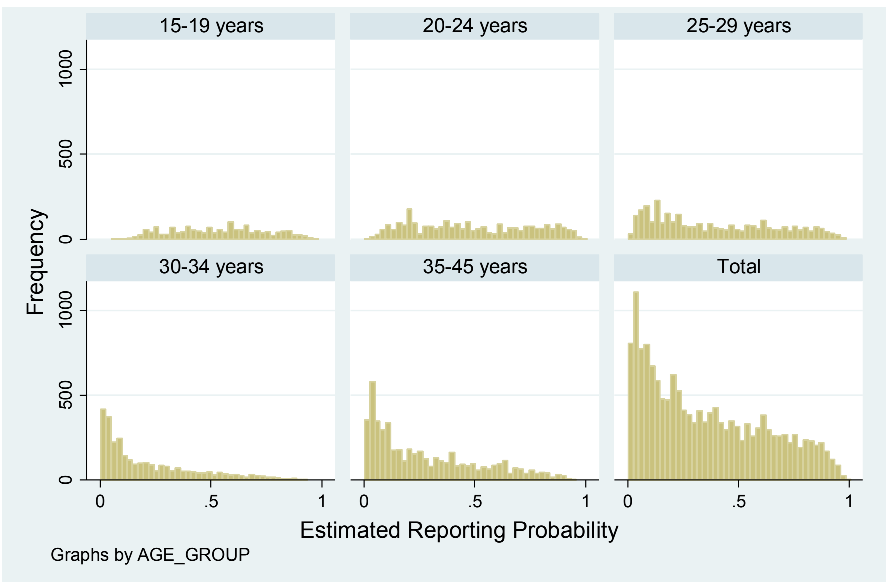

Table 6: Estimated reporting rates of different subgroups (%)

Group	Estimate	Std. Err. I		Std. Err. II	
All observations	35.28	8.23	***	15.23	***
Age group					
15–19 years	53.11	11.45	***	17.16	***
20–24 years	48.71	10.49	***	15.88	***
25–29 years	39.06	10.21	***	15.50	***
30–34 years	22.15	11.15	***	16.75	***
35–45 years	27.67	9.34	***	14.39	***
Race					
White	40.91	6.43	***	16.33	***
Black	20.48	10.55	***	11.80	***
Other	33.15	11.79	***	15.96	***
Hispanic origin	19.10	10.84	***	17.62	***
Education					
Below 12 years	17.59	11.18	***	16.79	***
12 years	22.75	10.26	***	15.56	***
Above 12 years	48.96	6.52	***	11.64	***
Religion					
No religion	37.16	10.26	***	15.57	***
Catholic	37.97	10.02	***	15.25	***
Protestant	28.27	7.48	***	12.29	***
Other	65.96	12.21	***	18.21	***
Marital status					
Never married	43.48	8.09	***	12.92	***
Married	32.17	8.58	***	13.47	***
Previously married	24.58	11.80	***	17.64	***
Cohabiting	31.00	11.39	***	17.08	***
Poverty level income					
below 100%	24.50	9.90	***	15.11	***
100–199%	29.08	10.21	***	15.50	***
200–299%	39.68	10.79	***	16.27	***
300–399%	45.63	11.16	***	16.77	***
400–499%	50.55	11.92	***	17.81	***
Above 500%	50.01	11.85	***	17.71	***

* p<0.1,**p<0.05,***p<0.01 against the null hypothesis that there is no underreporting, i.e., the reporting rate is 100%.
The standard errors were calculated using the delta method based on the ML parameter estimates. Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

The coefficient estimates of the variables that we included in the reporting equation are useful to identify the potential factors that would affect the reporting rate of abortions. However, they are not directly comparable to the unconditional average reporting rates of different subgroups estimated by previous researchers. The estimated unconditional average reporting rates that we report in Table 6 facilitate this comparison. As estimated in all previous research (Jones and Forest 1992; Udry et al. 1996; Fu et al. 1998; Jones and Kost 2007), Whites have a higher reporting rate of 40.9%, while the reporting rate of Blacks is only 20.5%. Other racial groups report 33.2% of their abortions. The reporting rate of Hispanics is only 19.1%. Unconditional means show that the reporting rate drops with age until 35 years but increases thereafter. The average reporting rate in the teen category (15–19 years) is as high as 53.1%. It drops gradually to reach 22.2% when the age is 30–34 years before picking up again when the age is 35–45 years.

As in Fu et al. (1998), the unconditional reporting rate increases with income even though the causal effect of income shows no clear direction according to our estimates. The reporting rate increases gradually from 24.5% for people below poverty income to over 50% for people having an income four times or more than the poverty income. Those who have less than 12 years of formal education report only 17.6% of their pregnancies according to our estimates, while that rate is 22.8% among the women with 12 years of education and as high as 49.0% among those who have more than 12 years of formal education.

The unconditional mean reporting rate among the married is 32.2%. The reporting rate is highest among those who have never been married (43.5%) while relatively low among those who have previously been married (24.6%). The reporting rate of cohabiting women (31.0%) is closer to the rate of married women. Catholics are more likely to report after controlling for other factors, but their unconditional reporting rate (38.0%) is only marginally higher than the rate of nonreligious women (37.2%). The reporting rate is 28.3% among Protestants and 66% among other religious categories. For each of these groups, the reporting rate is not significantly different from the overall average of 35.3%. We tested each of these estimates versus the null hypothesis that there is no underreporting. The null hypothesis is rejected with probabilities of over 0.99 for each of the subgroups, in addition to the entire group.

5.3 The robustness of results

In general, models proposed in the literature that attempt to deal with underreported counts are identified through nonlinearities in the log-likelihood function, and therefore they are usually sensitive to different specifications. The model we propose is not an exception unless there are one or more available variables that determine the reporting process but not the count process. Therefore, we did a series of robustness checks by changing specifications. The estimated average reporting rates using these alternative specifications are presented in Table 7.

Table 7: Estimated reporting rates from alternative specifications

Model	Reporting rate	Std. Err. 1	Std. Err. 2
Baseline model. Results are reported in Tables 3, 4, 5, and 6.	35.28	8.23	15.23
Alternative maximum of the baseline model.	28.58	7.71	12.32
Exposure period was counted from menarche ignoring the age at first sex.	35.67	8.03	15.21
The number of live births was included in the reporting process.	34.94	8.68	17.89
The two birth cohort dummies were included in the reporting process.	37.80	8.93	18.99
Both the number of livebirths and the two birth cohort dummies were included in the reporting process.	36.89	9.48	22.40
'Never worked' dummy was excluded from both count and zero-inflation processes.	35.95	8.42	15.30
Income categories were added as covariates of the count process.	54.33	8.30	21.38
Income categories were added as covariates of both the count and zero-inflation processes.	53.08	8.44	19.20
Marital-status categories were added as covariates of both the count and zero-inflation processes.	49.95	11.85	33.20
Income and marital-status categories were added as covariates of both the count and zero-inflation processes.	67.85	4.71	9.60
Income categories were excluded from the reporting process.	35.80	8.36	13.84
The number of formal marriages and the number of cohabiting unions were not included.	29.12	6.12	11.75
Survey weights were not used.	25.82	11.44	25.45
A negative binomial functional form was assumed.	27.68	7.47	12.14

The standard errors were calculated using the delta method based on the ML parameter estimates. Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

A more fundamental identification issue is distinguishing the global maximum of the log-likelihood function, since the result we have could correspond to the alternative local maximum. Therefore, we followed the suggestions in Papadopoulos and Santos Silva (2008) and used $[\beta + \gamma, -\gamma]$ as the initial values based on the estimated parameter vector $[\beta, \gamma]$ of our baseline model to find the alternative maximum. The coefficient γ of any variable included in the count process but not in the reporting process is assumed to be zero. The solution using these initial values results in a lower log-likelihood value, confirming that our baseline results, in fact, correspond to the global maximum. Moreover, according to this alternative solution, Whites are less likely to report abortions compared to other ethnic groups. The unconditional probabilities estimated using this alternative set of parameters suggest that the reporting rate of Whites is only 21.8%, while the reporting rates of Blacks and other races are 42.5% and 39.3%, respectively. This is contrary to the findings of all previous studies, which unambiguously show that the reporting rate is higher among Whites. The estimated reporting rate is 28.6% if we use this alternative set of parameters.

In our baseline model, the exposure period was counted from the menarche of a respondent or the age at first sex as reported. This window could be too short if some respondents have had sex before the date reported. Moreover, there's a possibility that some respondents are "fake virgins." In order to check the robustness of our estimates to measurement error in self-reported age at first sex, we reconstructed the exposure variable ignoring the self-reported age at first sex and assuming that every woman is exposed to abortions since her menarche. The resultant average reporting rate is 35.7%, which is not very different from our baseline estimate.

The number of live births a woman has had is a covariate of the count process in our baseline model, but we haven't included that variable as a covariate of the reporting process. Similarly, we have two birth cohort dummies among the covariates explaining the zero-inflation process, which are not among the covariates explaining the reporting process. What if we include these variables as covariates of the reporting process? When the number of live births is added as a covariate of the reporting process, the average reporting rate is 34.9%. Inclusion of the two birth cohort dummies increases it to 37.8%. When all three variables were added, the estimated reporting rate is 36.9%.

We have included marital status and income categories as covariates of the reporting process but not as covariates of the other two processes. The reason, as we explained above, is that only the current statuses of these variables are available. Unlike the changes in one's religion, changes in marital status and income are neither low-frequency events nor random events. During the exposure window for these women, on average, their income is likely to have increased, and those who were never married, a category that also excludes those who have been cohabiting, are likely to be in married, cohabiting, or previously married categories now.

We use the 'never worked' dummy as a covariate of the count and zero-inflation processes to identify the impacts of more-conservative cultural traits and the differences in preferred work–family balance on the number of abortions. The inclusion of this variable may potentially generate biases in the estimates of other variables if any of those variables are correlated with the 'never worked' dummy. To check this, we run our model excluding this dummy from all processes. Our baseline results are robust to this change, and the average reporting rate estimated using this model, 35.9%, is very close to our baseline estimate.

To explore the consequences of including the income and marital-status variables as covariates of the count process or the zero-inflation process, let's assume that the never-married and low-income women are more likely to have abortions. Consequently, if we include marital-status dummies as covariates of the counting or zero-inflation processes, we would overestimate the impact of married women on the number of true abortions. For the same reason, we would also overestimate the impact of income. As our model jointly determines reporting behavior and the abortion count, bias in the count process will also bias the estimates of the reporting probabilities. Alternative specifications show that the inclusion of these variables as covariates of the count and zero-inflation processes, in fact, would bias the estimated reporting probabilities upward. When the income categories are added to the count process, the estimated reporting rate increases by 19.1%. When income is added to both count and zero-inflation processes, it increases by 17.8% and jumps to 53.1%. Similarly, adding marital-status categories in those two processes causes the reporting rate to increase by 14.7%, and adding both income and marital status to the two processes in addition to the reporting process causes the estimated reporting rate to increase by 32.6% to 67.9%. This suggests that the income and marital-status categories should not be among the vectors of covariates explaining the count and zero-inflation processes. The results, however, are robust when income categories are excluded from the reporting process, and it indicates that the average reporting rate is 35.8%.

Even though we do not include current marital status as a covariate of the count process or the zero-inflation process, we include two other variables related to marital relationships: the number of formal marriages and the number of cohabiting unions of a respondent. These two variables show the stability of a woman's relationships during the entire period she has been sexually active, which is the same period she has been exposed to intentional abortions. Therefore, we believe that the inclusion of these two variables does not cause the same issues that we face when we include the current marital status or income. The results from a model excluding these variables show that a bias, if any, is likely to be in the opposite direction. When these variables are excluded, the estimated overall reporting rate decreases by 6.1%. As an additional robustness check, when we run our baseline model without using survey weights, the estimated average reporting rate is 25.8%, probably showing the effect of the lower reporting rates of Blacks and Hispanics (see Table 6) who are oversampled in this survey. Finally, we tested the model assuming a negative binomial functional form, which takes care of any residual over-dispersion not explained by the zero-inflation process. The estimated dispersion parameter (1.077 S.E. 0.108) is not statistically different from unity, suggesting that any residual over-dispersion not explained by our baseline model may not be a serious problem. The resultant average reporting rate, reported in Table 7, is 27.68%. Overall, these robustness checks suggest that the reporting rate should most likely be within the 28%–38% range.

6. Conclusions

In the United States, as a rule of thumb, less than one in two intentional abortions are reported in major surveys of women. Reporting rates in these surveys have typically been estimated by comparing with external data collected from service providers, which themselves underreport the correct number of abortions. Some researchers have compared with medical records and insurance claims, but these studies are based on small samples. In some other studies, the authors have attempted to use the discrepancy with a validation sample to identify the reporting rates. We demonstrate in this paper that the underreported counts of intentional abortions in surveys of women can be estimated with reasonable accuracy using a structural model. Not only does our approach eliminate the need for a gold standard or a validation sample, but it also helps to identify the factors causing underreporting.

The binomial-thinned Poisson model that we use here to explain underreported counts is not new. As we discussed before, the binomially-thinned zero-inflated negative binomial model proposed recently by Papadopoulos (2014) is very similar to ours. However, these models have been used very infrequently. Our results show the power of these models when used in an appropriate context. When count data in a survey is suspected to be underreported and there's no reason to believe that there is any overreporting, this model can effectively measure the extent of measurement error in counts, either with zero-inflation as here or without. When overreporting is suspected, a modified version of this model may be used in addition to several other count data models available.

Previous research shows that reporting rates vary significantly across different subgroups. From a policy perspective, it is important to identify any causal factors that affect the reporting rate so that the designers of a survey can attempt to address any vulnerabilities. Differences in reporting rates between various demographic groups, estimated by previous researchers through raw comparisons, are useful for this purpose, but they only provide some suggestive evidence about the causal factors that might influence the reporting probabilities of respondents. Using our parameter estimates, we can estimate the unconditional mean reporting rates of each individual and various subgroups. As we show using NSFG data, comparisons of raw means of different subgroups may lead to wrong impressions about causality. Even without a causal interpretation, the model can be used to predict the reporting probabilities of each individual, which in turn can be used to correct the bias due to measurement error in a model in which the number of abortions is used as the dependent variable or as one of the covariates. When estimated reporting rates are used in a model, however, the standard errors should be adjusted to account for the additional randomness, which we leave for future research.

In our model, the reporting probability does not depend on the true count of abortions, unlike in the models presented in Winkelmann (1998). We did not investigate how our results would be affected if we allowed the reporting rate to depend on the true count of abortions.

Our main intention here is to show that a researcher has an alternative tool to estimate the extent of underreporting in a given dataset, particularly when data from a reliable external source is not available for comparison. We did not, however, compare our Poisson model with the negative binomial model proposed by Papadopoulos (2014). That model is another alternative a researcher may want to consider in a setting similar to ours.

As usual with any estimate based on structural modeling, the estimates produced can only be as precise as the assumptions made in the model. If the true data-generating process is different from what we assume, the results can be biased. In our study, we only used the covariates available free of charge in public-use data files. In the restricted data files, there are additional variables, such as geographic information linked with measures of the availability of abortion services, which may be used to strengthen the identification of model parameters.

7. Acknowledgements

We are grateful for the valuable comments of two anonymous referees, the editor, Dean Lillard, Zhuang Hao, seminar participants of The Ohio State University, South Dakota State University, IUPUI, International Association of Applied Econometrics, and the Midwest Economic Association, whose comments and suggestions helped immensely to improve this paper.

References

Adimora, A.A., Schoenbach, V.J., and Doherty, I.A. (2007). Concurrent sexual partnerships among men in the United States. *American Journal of Public Health* 97(12): 2230–2237. doi:10.2105/AJPH.2006.099069.

Amoros, E., Martin, J.L., and Laumon, B. (2006). Underreporting of road crash casualties in France. *Accident Analysis and Prevention* 38(4): 627–635. doi:10.1016/j.aap.2005.11.006.

Cameron, A.C. and Trivedi, P.K. (2013). *Regression analysis of count data.* Second ed. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139013567.

Dvorzak, M. and Wagner, H. (2016). Sparse Bayesian modelling of underreported count data. *Statistical Modelling* 16(1): 24–46. doi:10.1177/1471082X15588398.

Fader, P.S. and Hardie, B.G.S. (2000). A note on modelling underreported Poisson counts. *Journal of Applied Statistics* 27(8): 953–964. doi:10.1080/02664760050173283.

Finer, B.L. and Zolna, M.R. (2014). Shifts in intended and unintended pregnancies in the United States, 2001–2008. *American Journal of Public Health* 104(S1): S43–S48. doi:10.2105/AJPH.2013.301416.

Fu, H., Darroch, J.E., Henshaw, S.K., and Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives* 30(3): 128–133, 138.

Goodkind, D.M. (2004). China's missing children: The 2000 census underreporting surprise. *Population Studies* 58(3): 281–295. doi:10.1080/0032472042000272348.

Goodkind, D.M. (2011). Child underreporting, fertility, and sex ratio imbalance in China. *Demography* 48(1): 291–316. doi:10.1007/s13524-010-0007-y.

Grindlay, K. and Grossman, D. (2013). Unintended pregnancy among active-duty women in the United States military, 2008. *Obstetrics & Gynecology* 121(2): 241–246. doi:10.1097/AOG.0b013e31827c616e.

Gross, A.M., Winslett, A., Roberts, M., and Gohm, C.L. (2006). An examination of sexual violence against college women. *Violence against Women* 12(3): 288–300. doi:10.1177/1077801205277358.

Henshaw, S.K. (1998). Abortion incidence and services in the United States, 1995–1996. *Family Planning Perspectives* 30(6): 263–270, 287.

Jagannathan, R. (2001). Relying on surveys to understand abortion behavior: Some cautionary evidence. *American Journal of Public Health* 91(11): 1825–1831. doi:10.2105/AJPH.91.11.1825.

Jones, E.F. and Forrest, J.D. (1992). Underreporting of abortion in surveys of U.S. women: 1976 to 1988. *Demography* 29(1): 113–126. doi:10.2307/2061366.

Jones, R.K. and Kost, K. (2007). Underreporting of induced and spontaneous abortion in the United States: An analysis of the 2002 National Survey of Family Growth. *Studies in Family Planning* 38(3): 187–197. doi:10.1111/j.1728-4465.2007.00130.x.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1): 1–14. doi:10.2307/1269547.

Lindberg, L.D. (2011). Unintended pregnancy among women in the US military. *Contraception* 84(3): 249–251. doi:10.1016/j.contraception.2011.01.017.

Papadopoulos, G. (2014). Immigration status and property crime: An application of estimators for underreported outcomes. *IZA Journal of Migration* 3(1): 12. doi:10.1186/2193-9039-3-12.

Papadopoulos, G. and Santos Silva, J.M.C. (2008). Identification issues in models for underreported counts. Colchester: University of Essex, Department of Economics (Discussion Paper Series; no. 657).

Papadopoulos, G. and Santos Silva, J.M.C. (2012). Identification issues in some double-index models for non-negative data. *Economics Letters* 117(1): 365–367. doi:10.1016/j.econlet.2012.06.001.

StataCorp (2015). *Stata statistical software: Release 14*. College Station: StataCorp LP.

Staub, K.E. and Winkelmann, R. (2013). Consistent estimation of zero-inflated count models. *Health Economics* 22(6): 673–686. doi:10.1002/hec.2844.

Tapales, A. and Finer, L. (2015). Unintended pregnancy and the changing demography of American women, 1987–2008. *Demographic Research* 33(45): 1257–1270. doi:10.4054/DemRes.2015.33.45.

Tennekoon, V. and Rosenman, R. (2014). Behold, a virgin is with HIV! Misreporting sexual behavior among infected adolescents. *Health Economics* 23(3): 345–358. doi:10.1002/hec.2918.

Udry, J.R., Gaughan, M., Schwingl, P.J., and van den Berg, G.J. (1996). A medical record linkage analysis of abortion underreporting. *Family Planning Perspectives* 28(5): 228–231. doi:10.2307/2135842.

US Department of Health and Human Services (2014). *Public Use Data File Documentation 2011–2013* [electronic resource]. Hyattsville: National Survey of Family Growth Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/data/nsfg/NSFG_2011-2013_UserGuide_MainText.pdf.

Winkelmann, R. (1998). Count data models with selectivity. *Econometric Reviews* 17(4): 339–359. doi:10.1080/07474939808800422.

Winkelmann, R. and Zimmermann, K.F. (1993). *Poisson-logistic regression*. München: Volkswirtschaftliche Fakultät der Ludwig-Maximilians-Universität München.

Yan, T., Kreuter, F., and Tourangeau, R. (2012). Latent class analysis of response inconsistencies across modes of data collection. *Social Science Research* 41(5): 1017–1027. doi:10.1016/j.ssresearch.2012.05.006.

Appendix

The following Stata/Mata code demonstrates how to estimate the zero-inflated binomial-thinned Poisson model through maximum likelihood estimation method.

```
    //Creating a simulated dataset
    clear
    set seed 1000
    set obs 15000
    gen t=trunc(20*runiform())+1
    gen x1=rnormal()
    gen x2=rnormal()
    gen x3=rnormal()
    gen x4=rnormal()
    gen pi=invlogit(0.3+x1-x2)
    gen lambda=exp(-1+x1-2*x3)
    gen e=logit(runiform())
    gen shi=(-0.6+x1-x3+e)>0
    gen pilamt=pi*lambda*t
    gen y1=rpoisson(pilamt)
    gen y=(1-shi)*y1

    //MLE procedure of the Zero Inflated Binomial Thinned
Poisson Model
    program drop urzip1
    program urzip1
    version 14
    args lnf theta1 theta2 theta3
    tempvar lambda psi
    quietly gen double `lambda' =
exp(`theta1')*invlogit(`theta2')
    quietly gen double `psi' = invlogit(`theta3')
    quietly replace `lnf' = ln(1-`psi')-
`lambda'+$ML_y1*ln(`lambda')-lnfactorial($ML_y1) if
$ML_y1>0
    quietly replace `lnf' = ln(`psi'+(1-`psi')*(exp(-
`lambda'))) if $ML_y1==0
    end

    //Estimation
```

```
    ml    model    lf    urzip1    (Count:y    =    x1    x2
x3,exposure(t))(Reporting:y= x1 x2 x3)(Positive:x1 x2 x3)
    ml max

    ml    model    lf    urzip1    (Count:y    =    x1    x2
x3,exposure(t))(Reporting:y= x1 x2 x3)(Positive:x1 x2 x3),
vce(robust)
    ml max
```

```
    //Average reporting rate
    mata
    Z=st_data(.,("x1","x2","x3"))
    b=st_matrix("e(b)")'
    gamahat=b[5..7]
    gamaC=b[8]*J(15000,1,1)
    Zgama=Z*gamahat+gamaC
    pihat=invlogit(Zgama)
    V=st_matrix("e(V)")'
    Vhat=V[5..8,5..8]
    Ggama=(Z'*(pihat:*(J(15000,1,1)-
pihat))\sum(pihat:*(J(15000,1,1)-pihat)))/15000
    Vpi=Ggama'*Vhat*Ggama
    mean(pihat)
    sqrt(Vpi)
    end
```

Table A-1: Simulation results

Variable	True Value	Estimate	Std. Err. I		Std. Err. II	
Count process						
X1	1.000	1.005	0.016	***	0.018	***
X2	0.000	-0.001	0.016		0.018	
X3	2.000	-2.005	0.010	***	0.010	***
Constant	1.000	-1.010	0.046	***	0.053	***
Ln(t)	1.000	1.000				
Reporting process						
X1	1.000	1.023	0.031	***	0.033	***
X2	1.000	-1.024	0.026	***	0.027	***
X3	0.000	0.005	0.027		0.026	
Constant	0.300	0.324	0.115	***	0.132	**
Zero-Inflation process						
X1	1.000	1.040	0.036	***	0.036	***
X2	0.000	0.007	0.026		0.026	
X3	-1.000	-1.080	0.039	***	0.039	***
Constant	-0.600	-0.669	0.038	***	0.038	***
Average reporting rate	55.538	55.921	2.010	***	2.310	***

* p<0.1,**p<0.05,***p<0.01 against the null hypothesis that the value is 0 for all coefficient estimates and against the null hypothesis that the reporting rate is 100% for the estimate of average reporting rate.

Std. Err. I is based on the empirical Hessian estimator and Std. Err. II is based on the Huber/White/sandwich estimator. The latter is robust to certain types of model misspecification.

The following Stata/Mata code was used to estimate the zero-inflated binomial-thinned negative binomial model through maximum likelihood estimation method.

```
   //MLE procedure of the Zero Inflated Binomial Thinned
Negative Binomial Model
   program drop urzinb
   program urzinb
   version 14
   args lnf theta1 theta2 theta3 theta4
   tempvar m p lambda psi
   qui gen double `m' = 1/ln(`theta4')
   qui gen double `lambda' =
exp(`theta1')*invlogit(`theta2')
   quietly gen double `psi' = invlogit(`theta3')
   qui gen double `p' = 1/(1+ln(`theta4')*`lambda')
   quietly replace `lnf' = ln(`psi'+(1-`psi')*`p'^`m') if
$ML_y1==0
   quietly replace `lnf' = ln(1-`psi')+
lngamma(`m'+$ML_y1) - lngamma($ML_y1+1) - lngamma(`m') +
`m'*ln(`p') + $ML_y1*ln(1-`p') if $ML_y1>0
   end
```